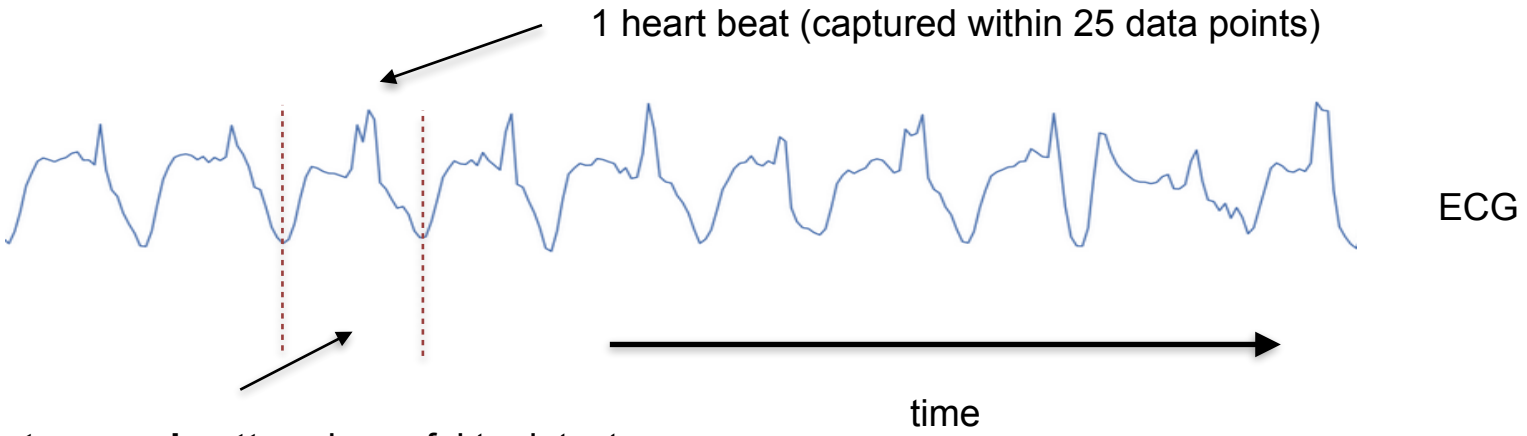
Window Size Selection In Unsupervised Time Series Analytics: A Review and Benchmark

AALTD'22, 23.09.2022, Grenoble, France

Arik Ermshaus, Patrick Schäfer, Ulf Leser

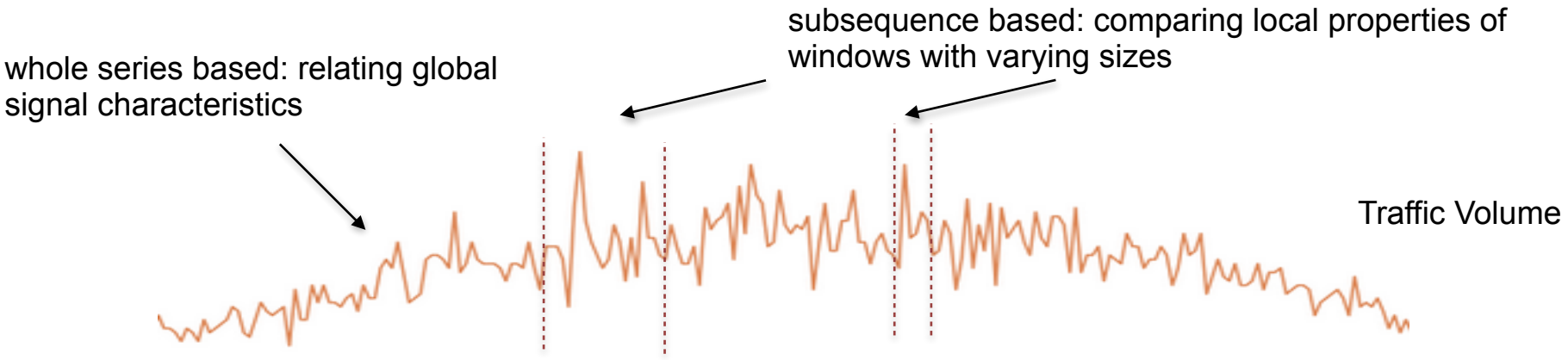
Temporal Patterns in Time Series



This **temporal pattern** is useful to detect semantic changes, anomalies or motifs in the signal

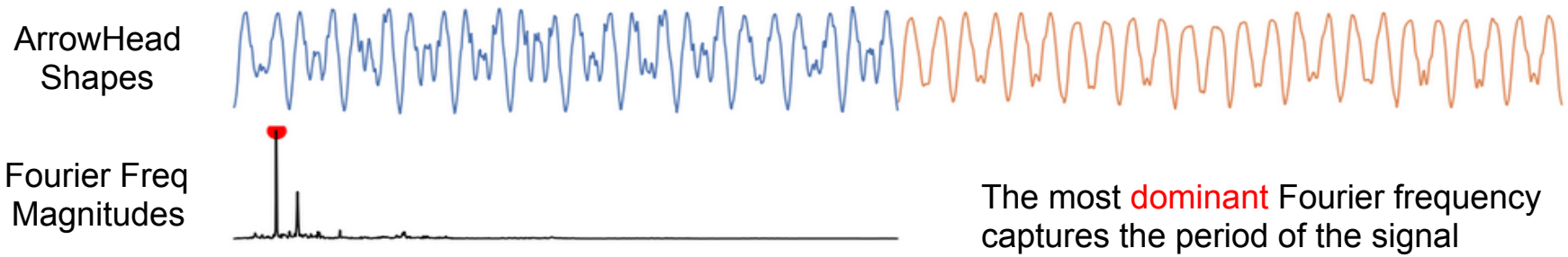
- Insights from TS can be drawn by inspecting local substructures
  - Temporal patterns often approximatively repeat throughout the signal
- Window size selection (WSS) is a preliminary task for many unsupervised TS analytics, e.g. EMMA [6], Matrix Profile [12], ClaSP [1], …
- Hence, determining the *right* window size of temporal patterns is a crucial task
  - TS analytics may have different requirements for window sizes

Window Size Selection: A Classification of Strategies



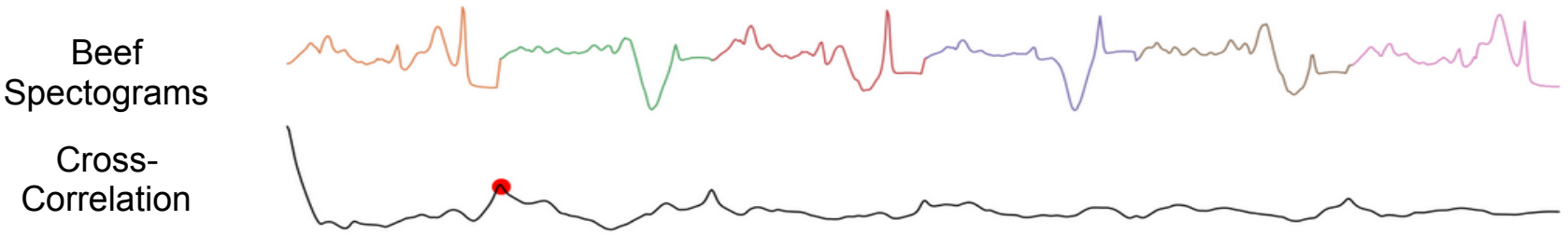
- Current WSS strategies try to detect dominant periods in signals
  - Assumption: TS approximately repeats a subsequence of values
- Whole series based techniques use frequency or time domain
  - Extract dominant frequency components or autocorrelated shifts
- Subsequence based methods extract local features from TS
  - Measure how well local window statistics align with global signal properties

Dominant Fourier Frequency (FFT)



- Fourier transform decomposes TS into sinusoid waves (Fourier coefs) which represent magnitudes of associated frequencies
- Whole series based method: Most dominant sinusoid wave (one with largest magnitude) captures a signal's period best
  - For each frequency: calculate magnitude
  - Return period length of most dominant Fourier frequency
- Dominant frequency is easy to extract, but can contain false positives
- Runs in $\mathcal{O}(n \log n)$ using $\mathcal{O}(n)$ additional space (n being the TS length)

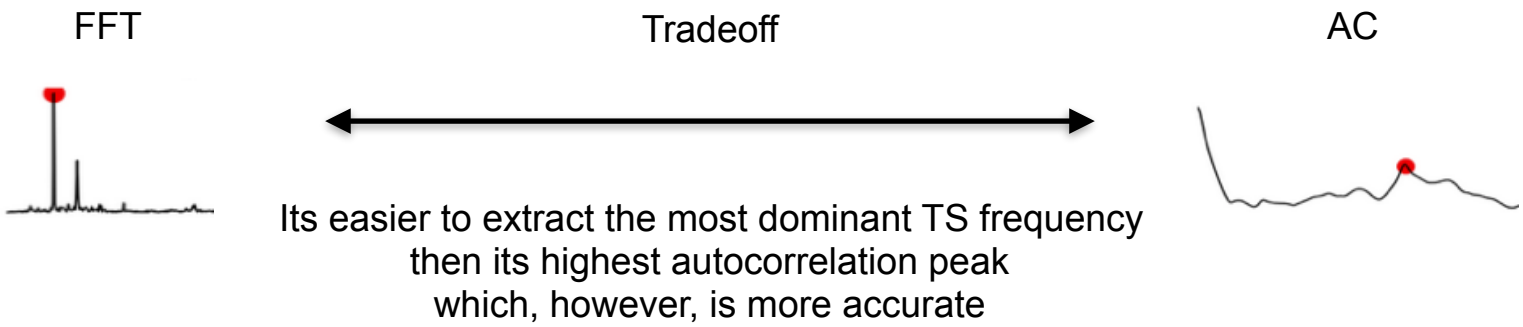
Highest Autocorrelation (ACF)



The AC has a substantial deflection that captures the period of the signal (within 118 data points)

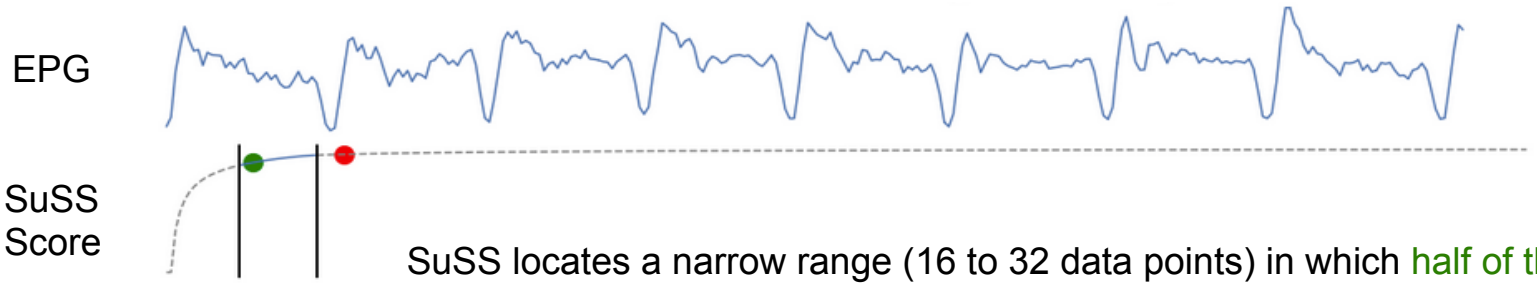
- Autocorrelation reports correlation of signal with delayed copy of itself for different shifts
- Whole series based method: Lag with highest cross-correlation captures a signal's period best
  - For each shift: calculate cross-correlation
  - Search for correlation with highest local maximum
  - Return the associated period length
- AC peaks are very accurate, but require a peak finder to be extracted
- Runs in $\mathcal{O}(n \log n)$ using $\mathcal{O}(n)$ additional space (n being the TS length)

Hybrids: AutoPeriod and RobustPeriod



- Hybrid methods try to overcome this tradeoff by combining both approaches
- AutoPeriod [10] computes a filtered Fourier transform and assigns dominant frequencies to AC hills to report the top location of the selected hills as a dominant period
- RobustPeriod [11] removes TS trend, decouples periodicities and detects dominant ones using modified variants of filtered Fourier transform and AC

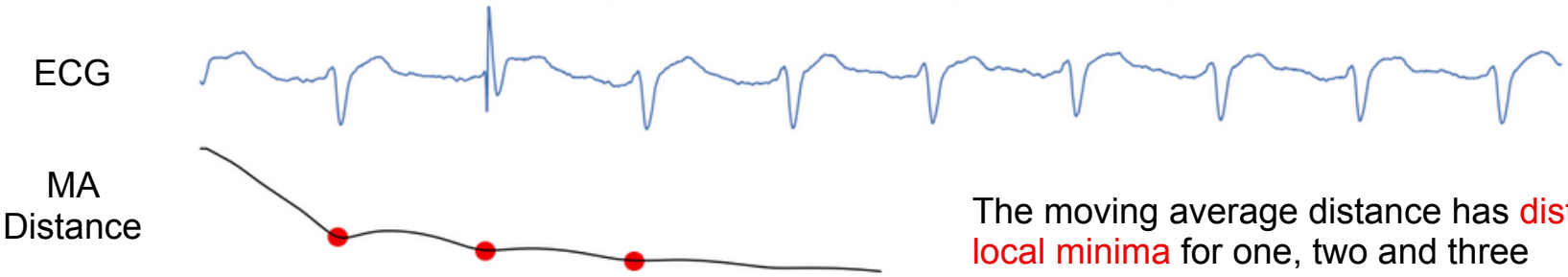
Summary Statistics Subsequence (SuSS)



SuSS locates a narrow range (16 to 32 data points) in which half of the period length (20 data points) is found

- SuSS [1] compares summary statistics computed over windows with the ones of the entire TS
- Subsequence based method: Summary statistics of appropriate window size are close to those of the whole signal
  - Perform exponential and binary search to locate window size with SuSS score larger then pre-defined threshold (fixed to 89%)
  - Return the 2x the window size as period length
- Very fast due to combination of two efficient search procedures
- Runs in $\mathcal{O}(n \log w)$ using $\mathcal{O}(n)$ additional space (n/w as TS/window size)

Multi-Window-Finder (MWF)



The moving average distance has distinctive local minima for one, two and three heart beat instances

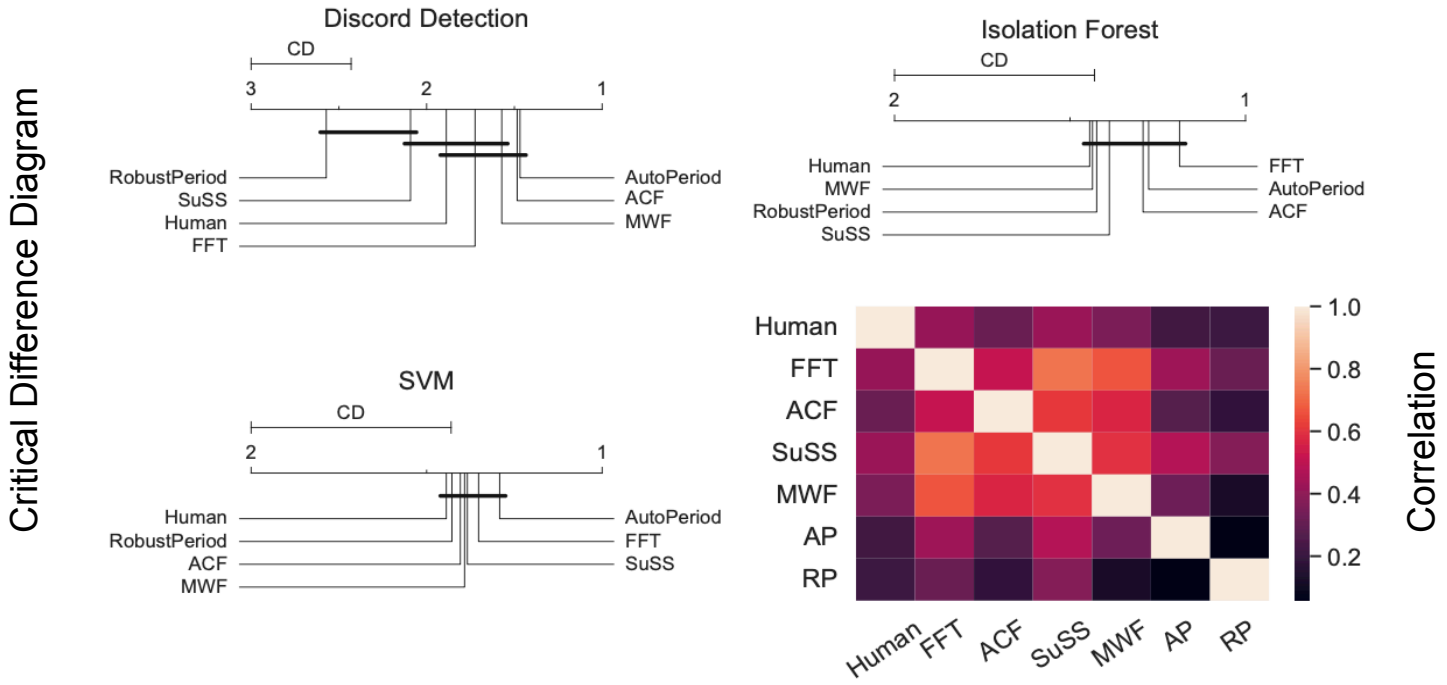
- Multi-Window-Finder [4] calculates moving average (MA) variances for a range of window size candidates
- Subsequence based method: Suitable window size has a small moving average variance
  - For each window size: calculate MA variance
  - Search for three smallest local minima in variances (using a peak finder)
  - Return their weighted mean position as period length
- Can be applied incrementally to extract multiple window sizes
- Runs in $\mathcal{O}(m \cdot n)$ using $\mathcal{O}(n)$ additional space (m window sizes, n as TS size)

Benchmark Setup

	Anomaly Detection	Segmentation	Motif Discovery
Data Sets	HEX UCR Anomaly Benchmark 2021 (HUAB) [5]; 250 TS each containing one anomaly	Time Series Segmentation Benchmark (TSSB) [1]; 83 TS with 1-9 segments	2 use cases (heartbeats [8], muscle activation [7])
Algorithms	Matrix Profile (MP) [12], Isolation Forest (IF), SVM	Window [9], FLOSS [2], ClaSP [1]	EMMA [6], Learning Motifs [3]
Metrics	F1 Score	F1 Score	Exploratory

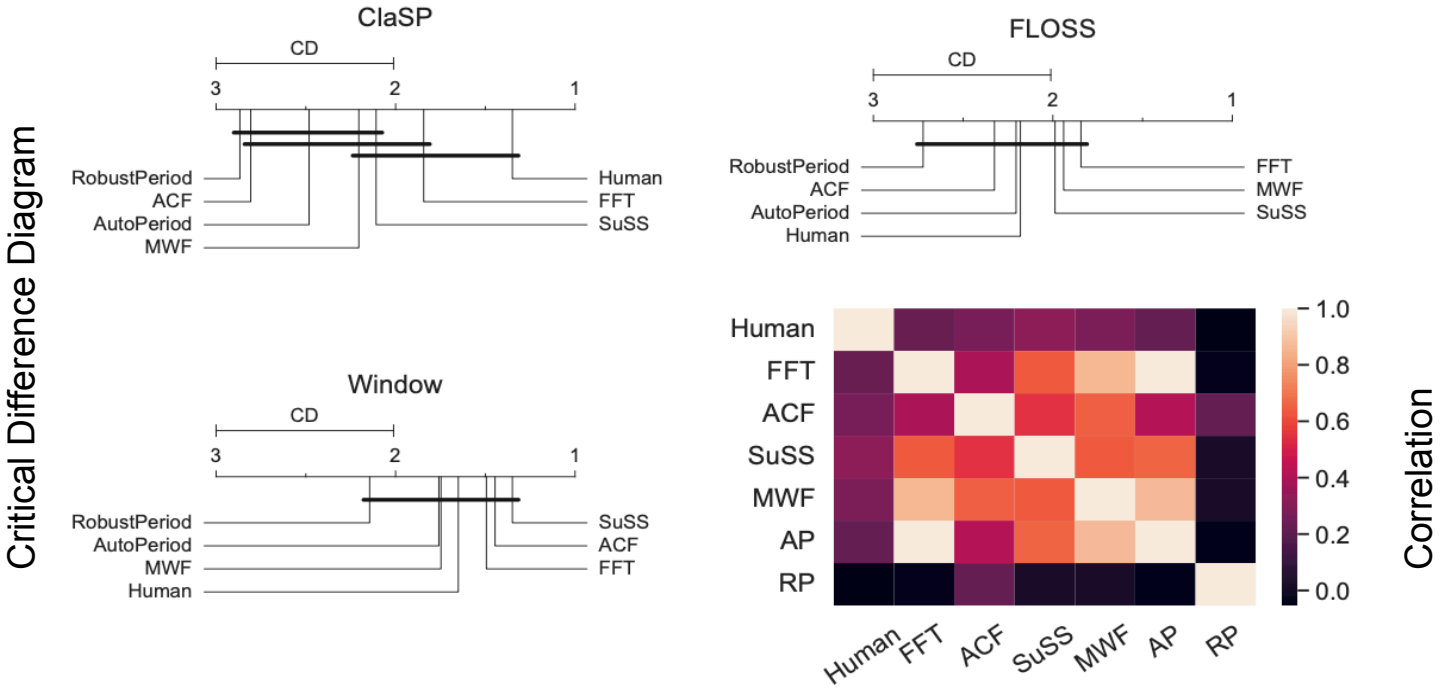
We tested each TS data mining algorithm with different window sizes on benchmark data sets and compare performances

Anomaly Detection



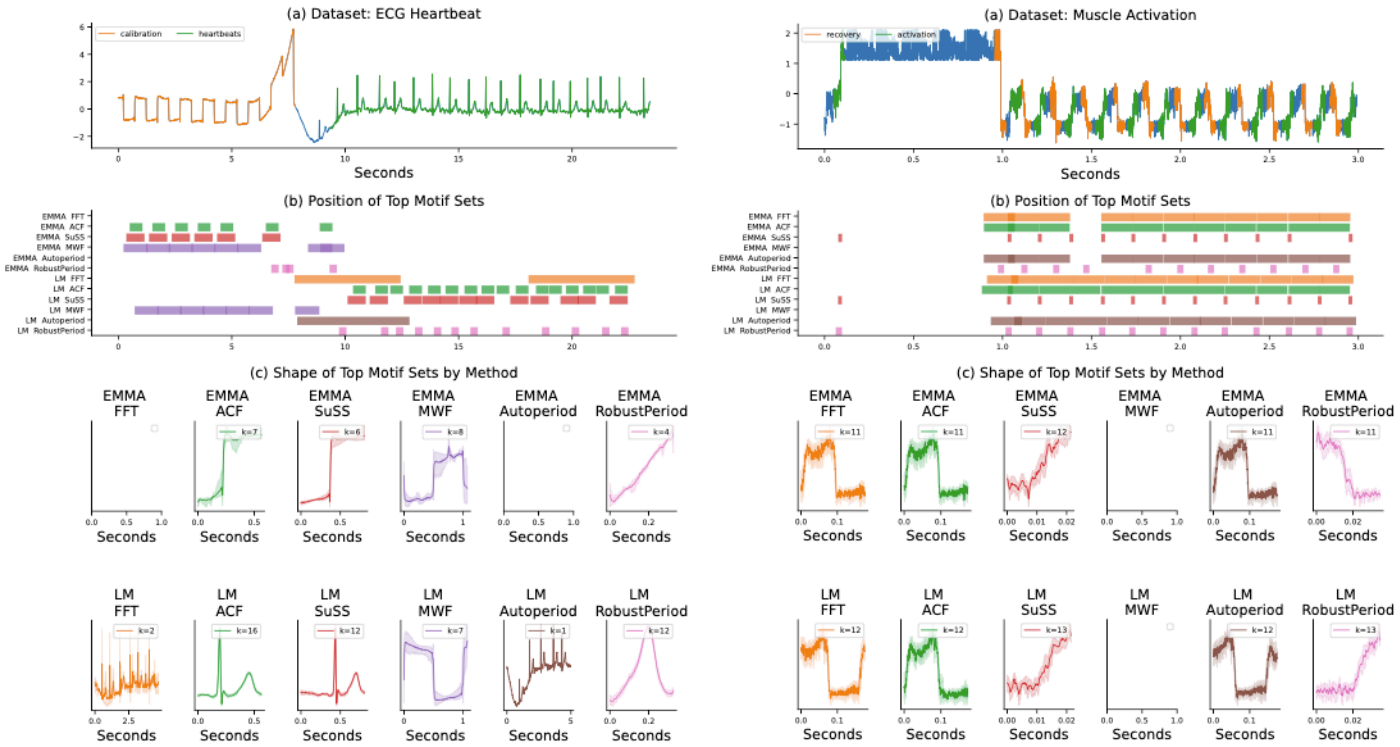
- Window size selection strategies perform comparably, but surprisingly better than human annotations
- AutoPeriod protrudes with weak correlation, but best-ranking results (hybrid approach may explain superior performance)

Segmentation



- Window size selection strategies are competitive with human annotations
  - Surprising, as TS have changing periods across segments
- FFT (whole series) and SuSS (subsequence) are top-ranking here, both may handle period changes better than AutoPeriod that performs worse

Motif Discovery



Window size selection strategies perform poorly, as TS do not have clear periods

Summary

- We applied WSS to unsupervised time series data mining tasks
  - Anomaly detection: the change of the period can even indicate the anomaly (AutoPeriod has superior performance)
  - Segmentation: each segment may have its own window size (FFT and SuSS can handle this best)
  - Motif discovery: TS must not be periodical, optimal window size may be independent of period (all methods gave unsatisfactory results)
  - Global ranking (across tasks): FFT, AutoPeriod / SuSS, ACF / MWF, human annotations, RobustPeriod
- Future work should consider …
  - non-periodic TS or ones with multiple dominant periods
  - Incremental / online detection of multiple window sizes
  - Testing more data mining algorithms on more data

References

1. Ermshaus, A., Schäfer, P., Leser, U.: ClaSP - Parameter-free Time Series Segmentation. arXiv (2022)
2. Gharghabi, S., Yeh, C.C.M., Ding, Y., Ding, W., Hibbing, P.R., LaMunion, S.R., Kaplan, A., Crouter, S.E., Keogh, E.J.: Domain agnostic online semantic segmentation for multi-dimensional time series. DMKD 33 (2018)
3. Grabocka, J., Schilling, N., Schmidt-Thieme, L.: Latent time-series motifs. TKDD 11(1) (2016)
4. Imani, S., Keogh, E.: Multi-window-finder: Domain agnostic window size for time series data. MileTS (2021)
5. Keogh, E., Dutta Roy, T., Naik, U. and Agrawal, A: Multi-dataset time-series anomaly detection competition. https://compete.hexagon-ml.com/practice/competition/39/ (2021)
6. Lonardi, J., Patel, P.: Finding motifs in time series. In: Workshop on Temporal Data Mining (2002)
7. Mörchen, F., Ultsch, A.: Efficient mining of understandable patterns from multivariate interval time series. DMKD 15(2) (2007)
8. Petrutiu, S., Sahakian, A.V., Swiryn, S.: Abrupt changes in fibrillatory wave characteristics at the termination of paroxysmal atrial fibrillation in humans. Europace 9(7) (2007)
9. Truong, C., Oudre, L., Vayatis, N.: Selective review of offline change point detection methods. Signal Processing (2019)
10. Vlachos, M., Yu, P.S., Castelli, V.: On periodicity detection and structural periodic similarity. In: SDM (2005)
11. Wen, Q., He, K., Sun, L., Zhang, Y., Ke, M., min Xu, H.: Robustperiod: Robust time-frequency mining for multiple periodicity detection. SIGMOD/PODS (2021)
12. Yeh, C.C.M., Zhu, Y., Ulanova, L., Begum, N., Ding, Y., Dau, H.A., Silva, D.F., Mueen, A.A., Keogh, E.J.: Matrix profile i: All pairs similarity joins for time series: A unifying view that includes motifs, discords and shapelets. ICDM (2016)

Code, Data and Experiments are available on
https://github.com/ermshaua/window-size-selection
Mail: ermshaua@informatik.hu-berlin.de